170-WP-008-001

Validation Issues and Processes for V0 Data Migration

White paper - Not intended for formal review or Government approval.

May 1996

Prepared Under Contract NAS5-60000

RESPONSIBLE ENGINEER

W. Waldron /s/, L. Klein /s/, M. Nishihama /s/, J. Pals /s/	5/28/96
Wayne Waldron, Larry Klein, Mash Nishihama, Jon Pals EOSDIS Core System Project	Date

SUBMITTED BY

Joy Colucci /s/	5/28/96
Joy Colucci EOSDIS Core System Project	Date

Hughes Information Technology Systems
Upper Marlboro, Maryland

This page intentionally left blank.

Abstract

This white paper examines issues and processes related to validation of data, and associated metadata, migrated from V0 to ECS standards. Some V0 data will be converted to HDF-EOS format, while other data will be left in native format. In all cases, ECS core metadata will be migrated and/or created from granule collection information in the DAAC inventory . We recommend that a validation plan be created for each V0 product, or similar product group, by the ECS V0 Data Migration Team in collaboration with DAAC personnel and data product specialists who are knowledgeable about the specific V0 product. This paper gives a framework from which to create such plans.

This page intentionally left blank.

Contents

1. Introduction

2. Data and Metadata Standards

3. Validation Issues and Processes

4. Validation Schemes

5. Recommendations

Appendix

References

1. Introduction

1.1 Purpose

Selected Version 0 (V0) data products will be migrated from their current formats to the Hierarchical Data Format - Earth Observing System (HDF-EOS) during the V0 Data Migration task. These data products will then be ingested into the Earth Observing System Data and Information System (EOSDIS) Core System (ECS). Validation of the migrated data is critical to ensure that the migration process accurately represents the original data. It is expected that DAAC scientists and other data specialists will be heavily involved in data validation.

The purpose of this white paper is to examine validation issues and processes related to V0 data migration and seek feedback from the V0 Distributed Active Archive Centers (DAACs) and science community. In particular, examination of various validation schemes must consider key drivers such as complexity, level of confidence, and cost. Although we expect that many of the methods discussed in this paper, at some level, will eventually be incorporated into V0 data migration, we stress that we are not restricted to these methods.

1.2 Organization

This paper is organized as follows:

1. Introduction
2. Data and Metadata Standards
3. Validation Issues and Processes
4. Validation Schemes
5. Recommendations

Appendix (Sample validation approach for an ERBE S-8 product)

References

1.3 Acknowledgments

Preparation of this White Paper was a team effort with input from many people. In particular, we would like to thank the ECS DAAC Science Liaisons for their review and useful comments. We would also like to thank, in advance, the DAACs and Science Community members who will review and offer their perspective and comments on the validation process.

1.4 Review and Approval

This White Paper is an informal document approved at the Office Manager level. It does not require formal Government review or approval; however, it is submitted with the intent that review and comments will be forthcoming.

Questions regarding technical information contained within this Paper should be addressed to the following contacts:

- ECS Contacts
  - Larry Klein, 301-925-0764, larry@eos.hitc.com
  - Jon Pals, 301-441-4320, jon@hops.stx.com
  - Tom Dopplick, 301-925-0333, tom@eos.hitc.com

Questions concerning distribution or control of this document should be addressed to:

Data Management Office
The ECS Project Office
Hughes Information Technology Systems
1616 McCormick Drive
Upper Marlboro, MD 20785

2. Data and Metadata Standards

2.1 V0 Standards

The EOSDIS V0 DAACs came into existence as separate, standalone data centers focusing on specific science expertise such atmospheric or oceanographic sciences. As a result, a great deal of diversity is to be expected across the DAACs in both data and metadata.

2.1.1 V0 Data Standards

There is no standard data format across the V0 DAACs. On the contrary, there is a plethora of data formats ranging from Hierarchical Data Format (HDF) to Committee on Earth Observations Satellites (CEOS) to principal investigator (PI) defined formats. For more details on the DAAC data holdings see the Science Data Plan [1]. Generally the V0 DAACs offer read software to account for this diversity of data formats.

2.1.2 V0 Metadata Standards

Metadata is data about data and initially the DAACs developed their own independent metadata models to support search and order for their data products. These independent metadata models continue to support the various local Information Management Systems (IMS) provided by the DAACs. With the coming together of the DAACs to form EOSDIS V0, a system level V0 IMS was formulated to allow inventory queries across all DAACs. This required the specification of a minimum set of metadata standards that could be implemented by all DAACs. Thus, a search via the local IMS may locate DAAC data not found via the system-level IMS.

2.2 ECS Standards

2.2.1 ECS Data Standard

ECS has baselined HDF-EOS as its standard data format in order to achieve the following goals:

- Make the data self-describing
- Make the data more easily accessible
- Standardize information classes
- Provide instrument-independent services
- Provide standard interface for passing data

HDF-EOS is HDF with defined conventions contained in files that are readable by standard HDF libraries. HDF-EOS allows both non-geolocated structures (N-dimensional array, Science Data Table, Raster Image, Text, and Inventory and Product-Specific Metadata) and Geolocated Structures (Point, Swath, and Grid). For more information on HDF-EOS see the HDF-EOS Primer for Version 1 EOSDIS [2].

The conversion of native data to HDF-EOS will require special care to ensure that all parameters and science data are correctly mapped to the HDF-EOS format. The original data may have a built-in logical relationship among the data elements and this relationship must be maintained in the converted HDF-EOS product. In particular, the relationship between data values and geolocation information is maintained by a concept of "structure" for point, grid or swath data in HDF-EOS. Information on these ideas and the "linked field" concept can be found in references [3], [4], [5], and [6]. Other relationships may be more difficult to define and validate. DAAC scientists can greatly help in this area.

2.2.2 ECS Metadata Standard

Metadata will be created and applied in ECS for various categories, including collection-level metadata (collections of data granules) and granule-level metadata (individual granules). In ECS nomenclature, collection-level and granule-level metadata are known as inventory metadata and will be stored both in the ECS inventory database and in the data product granules themselves. Inventory metadata applicable to all data products is provided in reference [7]. Further, ECS also allows the inclusion of product-specific metadata, which is specialized to individual granules. V0 data migration will migrate metadata as well as the data into ECS. This means acquiring and mapping V0 metadata into the ECS data model standard.

3. Validation Issues and Processes

3.1 Validation Issues

3.1.1 Data Validation Issues

The fundamental issue concerning data migration is the preservation of the integrity of the original data. Possible loss of numerical integrity can occur when the native and the migrated data sets reside on different system/hardware platforms. Further, data could be lost or corrupted during reformatting and reorganization if quality checking and validation procedures are not carefully developed and applied throughout the migration process. The proposed migration of V0 data products from native format to HDF-EOS format, along with their associated metadata, can be considered data reorganization and data reformatting.

Scientists are rightly concerned when data are reorganized or reformatted since any changes to the original data provide an opportunity for errors to enter the migrated data stream. Therefore, in order to obtain support from the scientists for data that is reorganized or reformatted, any proposed validation scheme must provide the highest level of confidence that the migrated data will be identical except for numerical differences caused by differences in hardware platforms. Numerical differences must be understood and agreement reached on the acceptable tolerances (e.g., 1 part in 10^7), when compared with the original data.

Ensuring a high level of confidence is not an easy task, and validation will be a significant cost driver in the data migration effort. However, the effort can be reduced by application of available COTS packages as well as collecting and applying experiences from other groups that have done similar migration tasks. For example, several DAACs have migrated data from native format to HDF format (e.g., Pathfinder Project). The High Energy Astrophysics Science Archive Research Center (HEASARC) at GSFC has spent the last several years generating and migrating to a common format (FITS) X-ray data base. In addition, commercially based data have also been migrated into common data bases. For example, the oil companies combined their efforts to put the large amount of geological data into a common data base.

Data validation and quality checking are integral parts of our end-to-end data migration process beginning with the development of conversion software and ending with the actual migration of the data at the DAACs. Our approach will apply rigorous software development techniques and reuse available validation tools, such as existing DAAC validation schemes, if available. Software development and validation are related through the use of "software metrics", where validation can be viewed as one of the requirements for system testing (refer to Grady & Caswell [8] or Jain [9]). Quality checking uses both embedded and system tools to ensure that the data migration process has been correctly executed and completed in migrating data from V0 to ECS. Approaches for data validation and quality checking will be developed jointly with the DAAC staffs and documented in Data Group Data Migration Plans. A data group can be a single data product or a group of data products with similar characteristics.

3.1.2 Metadata Validation Issues

The V0 migration effort will be faced with metadata written in many formats, in machine-readable form and some with hardcopy only. When compared with ECS mandatory metadata, we will find missing values and it will be necessary to create and validate them as part of the migration process.

We expect to encounter significant diversity among the various data sets and associated metadata. Since each data set can have its own set of uniquely defined instrument dependent metadata, there can be a variance in the definition of a given parameter among the instrument teams. For example, sea surface temperature may have a different meaning between atmospheric and oceanographic researchers.

Another issue will be diversity in storage methods. For example, HDF provides a well-defined method of writing metadata, but for non-HDF datasets, each data set will have to be inspected to determine the format. Additionally, metadata may exist in a file separate from the data file. It may exist in the general form (e.g., PARAMETER = VALUE), as a list of values, or as an ASCII text block. Presumably, the associated read software will specify the format and definition, but it will require a hands-on inspection in order to extract the information. Although these two points are more closely related to the general data migration issue, rather than a validation issue, it is important to realize that metadata validation goes beyond simply comparing values.

In principle, once the metadata have been mapped to the ECS standard, any approved validation scheme for alphanumerical data will also be valid for the metadata. The metadata can be inspected visually or electronically using COTS comparing schemes to look for discrepancies. For example, the EOSView utility is capable of allowing visual comparison of metadata values. NCSA's Vshow and hdp command line utilities will also provide useful tools to evaluate metadata values.

3.2 Validation Processes

3.2.1 ECS Pilot Data Migration Study

Validation issues were initially addressed by the ECS project in the Pilot Data Migration Study and documented in the final report [10]. The study used two basic validation methods, volume and data value examination. The pilot study focused on trying to determine factors that caused the discrepancies in the volumes. It was found that volume cannot be used as a definitive validation criteria due to large differences in native and migrated data set volumes. Data written as an HDF file, for example, may have a larger or smaller volume than the original binary format. Since the HDF-EOS conversion process physically reformats a data product, volume variability is to be expected.

Various data value examination methods were used in the pilot study, such as precision checking, displaying and checking data visually, and electronically comparing images. Specifically, the NCSA Vshow utility and the UNIX cmp program were used for data examination.

3.2.2 V0 Data Migration

For V0 data migration, we are exploring additional validation tools and adding more structure to the data validation process. In Section 4, we examine and explore possible validation schemes, methods, and approaches that could be used during V0 data migration. In Section 5 we offer recommendations and solicit feedback from the DAACs and science community on their perceptions of the data and metadata validation process.

This page intentionally left blank.

4. Validation Schemes

4.1 Data Validation Schemes

This section focuses on validation of data converted from native format to HDF-EOS because such conversions are a major migration challenge. Note, however, that HDF-EOS includes internal metadata stored as metadata objects, so validation of HDF-EOS data also involves schemes to validate the internal metadata. Data validation schemes can be divided into two general categories, each with several methods:

sampling schemes

- random verification method
    - -- individual data characteristics
    - -- visual data element check
    - -- numerical data element comparison
    - -- graphical data element comparison
- partial verification method

full verification schemes

- complete verification method
    - -- individual data characteristics
    - -- visual data element check
    - -- numerical data element comparison
    - -- graphical data element comparison
- reverse engineering method
    - -- software inversion
    - -- bit-to-bit comparison

This hierarchy is not complete and one could devise additional variations to the schemes and methods presented above. However, this hierarchy represents a reasonable starting point and we anticipate that an acceptable validation scheme for each migrated product will emerge as a collection of these core processes.

Compared to full verification, sampling gives less confidence in the final products. In either case, the logical relationship among the data elements must be checked, if applicable for the product, as emphasized before.

In addition, all validation schemes will have to specify a level of acceptable precision. Precision could be a potential problem based on the fact that many data sets were generated on older platforms and will be migrated using newer platforms. The precision problem will be inherent in all validation schemes, and allowances must be established before a decision on a validation scheme is finalized.

For data that are converted to HDF-EOS, an additional validation step can be performed in Release B by invoking ECS data services such as subsetting and subsampling. V0 offers no system-level subsetting services so any comparisons must be performed against local DAAC subsetting services, which are only available for a limited number of V0 products. If no V0 comparisons are available, then ECS data services must be evaluated on a case-by-case basis to verify that ECS data services can be successfully invoked using the data migrated and converted to HDF-EOS.

4.1.1 Sampling Verification Scheme

4.1.1.1 Random Verification Method

Random verification provides the least complex and least expensive validation method. However, it also provides the lowest level of confidence due to its random nature. This procedure can be used to compare reformatted and native data using one or all of the following four approaches:

1) Individual data characteristics

2) Visual data element check

3) Numerical data element comparison

4) Graphical data element comparison

The major disadvantage of random verification is that only a fraction of the data set is actually verified, with the assumption that the remaining data elements have migrated correctly.

In approach (1), a value will be calculated from a set of parameters in both native and migrated data sets individually. Then these two values are compared. Note that both data sets do not have to be on one system at the same time. A checksum is a good example of this technique. In approaches (3) and (4), two sets of the same parameters from the native and the migrated data can be on the same system and numerical differences could be used for comparison.

One strategy is to apply the random verification scheme after the full verification scheme is completed for a few sample granules of migrated data.

It should be noted that the term "random verification" is used in a fashion similar to the statistical term "random sampling". However, the reader should be very cautious about the use of randomness. In statistics, a subset (sample) of a population is selected randomly and parameters are estimated for the population. For a random sampling technique (such as during a manufacturing process) system parameters are computed and product reliability is evaluated with some confidence level. In some sense, our efforts on random verification are similar to a product test.

Another similarity would be that as more validation is performed, the confidence level will increase. This is similar to the idea of increasing sampling points in statistics. Because of this similarity, we use the term "random verification". For statistical sampling techniques, product reliability tests and other useful information, refer to Jain [9], Freund and Walpole [11], Hansen, Hurwitz & Madow [12], and Chorafas [13].

Individual Data Characteristics

Usually, the checksum technique is used in the data transfer and communication field to check if transferred data is corrupted. Sometimes data is transferred to a mass storage area and, at a later time, a checksum can be used to make sure that the data retrieved is not corrupted in any way. Detailed information and a related program can be found in [14].

The technique is not restricted to the communication field or to data storage problems. Volume checking can be useful for detecting obvious problems such as in those cases where the volume exhibits sudden, drastic changes during migration that indicates the probability of formatting errors. Also, a set of parameters can be manipulated and compared in two data sets (e.g., sum of all the elements in an array or image pixel values). An advantage of this method is that only a portion of parameters need be used for data validation.

Another approach is to check values for the migrated data set using known tolerances or valid ranges for attributes, such as ranges of temperature, air pressure, or radiances at certain geographical locations and altitudes. An automated process could generate an output file and/or generate a flag that would stop the migration when the acceptance criteria are not met. For example, by considering the inspection of the axes parameters (spatial and time), several obvious questions can be asked; 1) do the latitude and longitude variables increase or decrease in the right direction; 2) is the sign of the latitude correct, and; 3) is time chronological and positive with respect to a reference time? This could easily be done as an automated differencing scheme which could be checked either electronically or visually.

Visual Data Element Check

Metadata and Science data for native and migrated data sets can be checked on the screen or via hardcopy through the use of EOSView, Collage, Vshow and hdp command line utilities. This approach provides a quick check of parameters and is quite effective for a limited number of values. If both sets of parameters are displayed on screen at the same time, their comparison would be easily done with some accuracy.

Numerical Data Element Comparison

Sets of parameters in the native and migrated data sets can be read and their differences calculated. Any non-zero differences are checked to determine if the differences are acceptable machine differences. If two sets of parameters cannot easily be read by one program, they could be extracted to separate files and then compared. This example was given in the pilot report [10].

Graphical Data Element Comparison

Since graphical displays are a major step up from individual random sampling, due to their ability to compare a large amount of data during one display, the level of confidence is significantly increased. This can be done for both floating point and image data. The procedure is similar to the one described above in *Individual Data Characteristics*, except in this case, we choose a random set of numerical arrays and/or image arrays. The best way to validate a numerical array and/or image is to visually display the difference between the original array and the migrated array (although precision problems may arise due to machine dependent round-off errors). In a few limited cases, visual inspection may be sufficient. This can also be automated by a routine that takes these differences and produces a flag when the results are larger than a defined threshold. However, for both numerical and image comparisons, it may also be wise to visually inspect each plot or image periodically to ensure that both the original and migrated plots and images are non zero.

EOSView and NCSA's Collage will be useful for comparing two images. If native data is not in HDF format or images cannot be handled by the above utility programs, new images may be created in an HDF format or in a standard image format, which can then be displayed and compared by a utility program. A difference image may be created in this fashion. Image data comparisons were discussed in the pilot study.

4.1.1.2 Partial Verification Method

Partial verification is the method of examining the same region of data in each and every data granule. Consider the following example. Suppose the data set contains a large number of two dimensional arrays. We extract a specified number of columns and/or rows from each array and compare each element of the columns or rows with the associated original columns and rows. Another approach is to sum the columns or rows and make comparisons with the associated original sums. These two procedures can also be combined to increase the level of confidence. Furthermore, the level of confidence can be extended by increasing the number of rows and columns used. It is best to consider differences, as discussed above, in elements and sums. Although this procedure does test each data array, which is clearly advantageous over the random method, it does not test every element of the array. This is where the statistical nature arises. The level of confidence will be related to a probabilistic formula determined by the statistical sampling of rows or columns. The major disadvantage of this method is that we are considering sums and not individual data values, and again, not all data elements may be sampled.

4.1.2 Full Verification Scheme

Full verification can theoretically provide total assurance that the migrated data set is identical to the native data set (or to within a pre-determined tolerance). However, full verification also requires the most complex and expensive methods due to the degree of software development and level of computation time required.

One method that can be used for full verification is complete verification where every element is checked. A second method is reverse engineering where the migration process is reversed and the migrated data set is transformed back into its original state. However, this approach may not produce an exact copy of the original without a considerable cost.

4.1.2.1 Complete Verification Method

This method eliminates the uncertainties associated with random or partial verification by verifying that all comparable array elements in the native and migrated data sets are the same. By definition, since each migrated data element is compared with each original data element, the process provides total assurance of a successful migration. If additional information has been added beyond that provided in the original data, this additional data must also be completely checked. One or more of the approaches employed for random and partial verification (see Sections 4.1.1.1 and 4.1.1.2) can also be applied to complete verification.

4.1.2.2 Reverse Verification Method

In this method the migrated data set is converted back to its native format, and this "newly" created data set can be checked bit-to-bit with the original dataset. The data sets must be identical or at least identical to some level of agreed upon precision. If the volumes between the reverse migrated and native data sets are the same, then the only differences will occur due to precision differences from different platforms.

As an illustration, a reverse verification method could involve the following steps:

a. define a series of data inspection procedures for the original data (e.g., display of certain images, plots of different variables, etc.),

b. migrate the original data to the new format,

c. reverse the process and migrate the new formatted data back to the original data format,

d. verify that the "recreated" data set is identical in volume to the original data,

e. use the set of data inspection procedures on the "recreated" original data, and

f. if comparisons are identical to an agreed upon precision, then the migration has been successful

The main advantage of this method over the complete verification method is that it is not necessary to compare each element; volume and selective data inspection become the validation criteria. However, if there is no loss of precision, a full bit-to-bit comparison after reverse migration would also establish that the native and migrated data sets are identical.

4.2 Metadata Validation Schemes

Validation of metadata within HDF-EOS files was discussed in Section 3.1.2. Here we address the validation process for inventory metadata, i.e. the collection-level and granule-level metadata. The ECS standard for metadata is reference [7] and we know from the Pilot Data Migration Study [10] that some V0 metadata descriptions do not directly map to ECS metadata descriptions. Further, after the mapping between V0 and ECS metadata is completed, and values assigned to ECS metadata, we expect there will be missing ECS metadata since the ECS metadata model is more information rich than the V0 metadata models.

4.2.1 Collection-Level Validation Scheme

ECS Collection-level metadata provides a high-level description of a collection of data granules, or a group of collections. Preparation of collection-level metadata begins with the mapping of V0 and ECS metadata on a product-by-product basis. After the mappings are understood, values can be assigned to the equivalent ECS metadata. Validation of the population of ECS collection-level metadata is accomplished by iterative review with the V0 DAACs. Missing ECS metadata attributes are filled in through iterations with the DAACs and data providers, as appropriate. Validation of these additional metadata is inherent in the DAAC review process. ECS collection-level metadata are prepared before the actual migration of a collection begins at a DAAC.

After collection-level metadata are delivered and inserted into an ECS Data Server, they will be further examined using tools developed for maintenance and operations (M&O). Additionally, collection-level metadata can be viewed using the ECS Client and ECS services invoked based on collection-level information. Where applicable, comparisons can be made with equivalent Version 0 IMS services.

4.2.2 Granule-Level Validation Scheme

Granule-level metadata describe each data granule and become part of the searchable ECS inventory. Preparation of ECS granule-level metadata begins with the mapping of V0 and ECS granule-level metadata on a product-by-product basis. However, metadata values can be assigned only during the actual migration since the values will vary for each granule.

Validation of the V0-to-ECS mappings is accomplished by iterative review with the DAACs. During the actual migration at a DAAC, granule-level metadata are mapped from V0 to ECS, and then range checked upon delivery and ingest into ECS. After the inventory schema are populated with the granule-level metadata, a final check will be performed by creating the same search using the V0 and ECS search clients and the results compared to ensure the same granules can be found in both V0 and ECS.

5. Recommendations

The main driver in validation is the preservation of the integrity of the original data. However, ensuring a high level of confidence is not an easy task, and validation will be a significant cost driver in the data migration effort.

We recommend that the validation process be assembled from various combinations of validation schemes and methods on a product-by-product basis. There is great diversity in V0 data products; some products contain imagery, some products contain vertical profiles, some products contain solar observations, and the remainder contain a wide variety of other data types. The migration process needs to identify the best validation scheme for each product based on the recommendations and collaboration of people knowledgeable about the data product, i.e., data producers, DAAC User Services, and ECS data migration team. Further, any proposed validation plan needs to receive full review by the collaborators before implementing the validation process.

Full verification, although expensive and time consuming, should seriously be considered since this scheme has the ability to provide total assurance that the migrated data is identical to the original data. However, cost will be a significant factor. Initially, there should be a period where all data is rigorously checked but, as confidence grows in the migration process, sampling may become the principal validation scheme. Various combinations of the random and partial verification methods can then be applied depending on the granule size and complexity.

This page intentionally left blank.

APPENDIX

A Sample Validation Scheme for ERBE S-8 Granules

We have written a program which migrates the ERBE S-8 cloud product to an HDF-EOS swath structure and the metadata to ECS format. Validation of the ERBE S-8 Product migration can be accomplished by a combination of several methods discussed in this White Paper. The data may be verified graphically and with an element-by-element check.

The migrated HDF-EOS ERBE S-8 swath granule can be validated with graphical displays. The native and HDF-EOS fields can be plotted separately as shown in Figures A-1 and A-2 for the first four scans of the ERBE S-8 longwave data and then visually compared to see if there are differences.

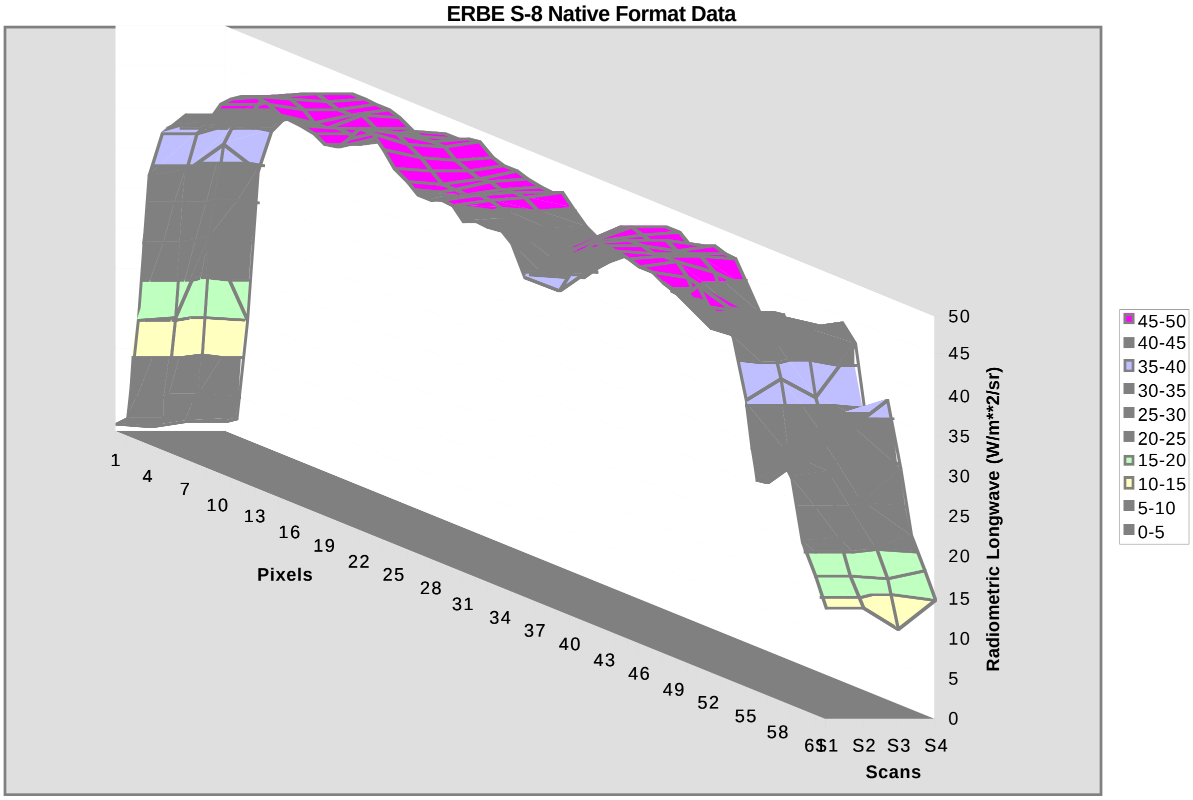

Figure A-1. ERBE S-8 Native Format

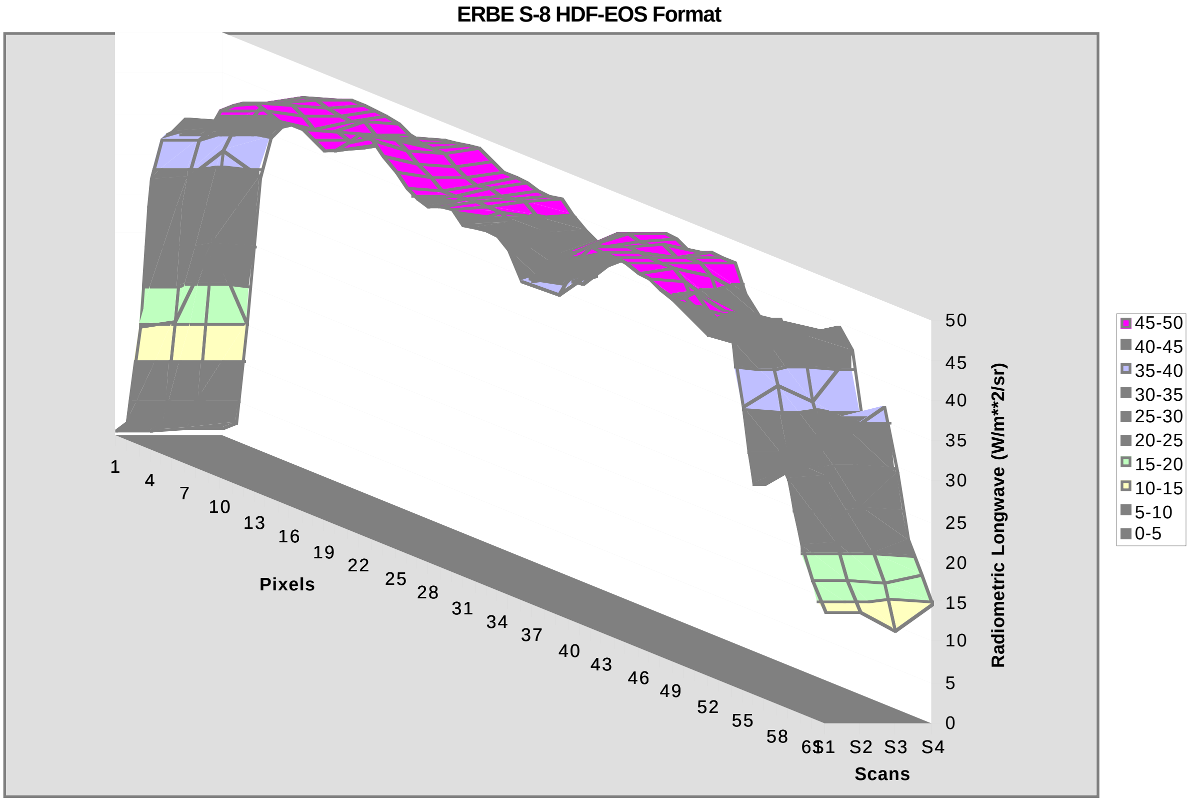

Figure A-2. ERBE S-8 HDF-EOS Format

The plots can be false color images (8 bit or 24 bit raster image), or the data may be gridded so that contour plots can be produced (possibly color filled).

An alternative is to compute the differences between the native and converted data and plot the differences. An example is shown in Figure A-3 for the differences between the ERBE S-8 native and HDF-EOS data after introduction of simulated small errors. This technique demonstrates the power of differencing as a tool for visual validation.

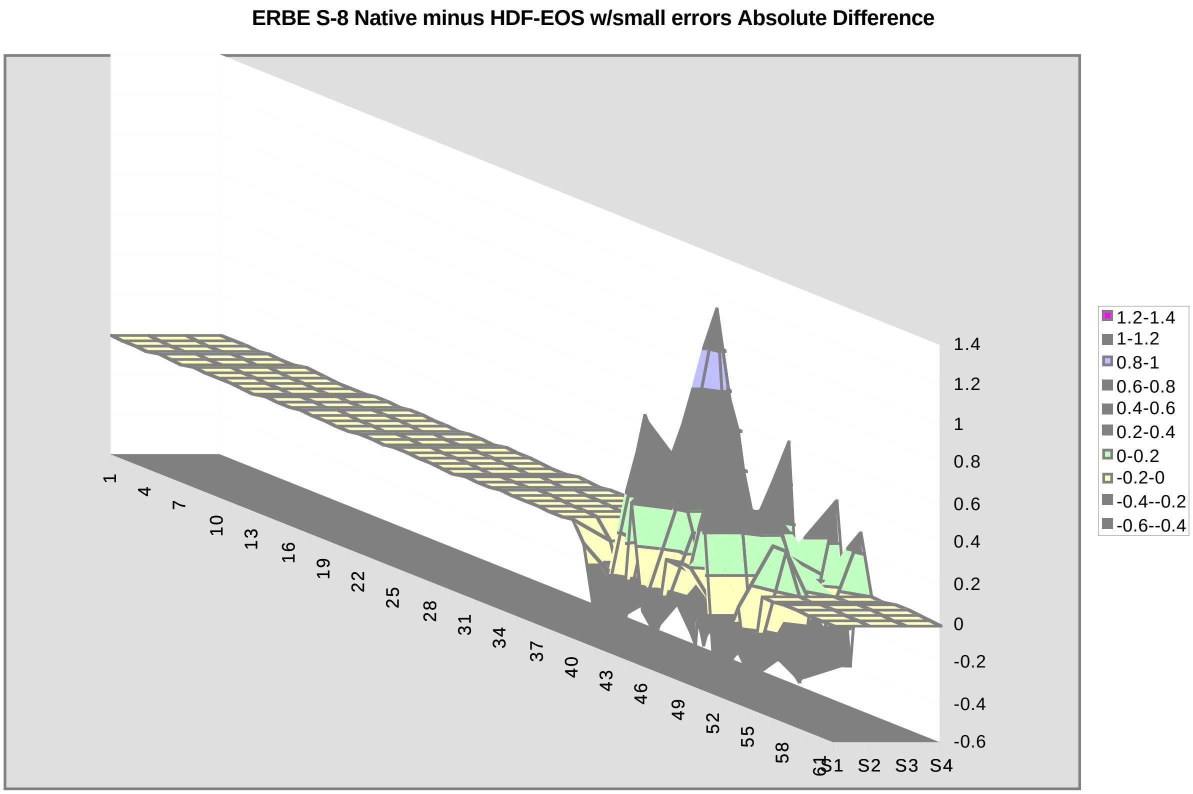

Figure A-3. Differences between ERBE S-8 Native and HDF-EOS After Introduction of simulated errors

The graphical analysis could also be used to verify geolocation fields (latitude, longitude, etc.), which should be consistent for each parameter field.

An element-by-element check of the data may also be performed. A binary element flag would be set whenever there is a discrepancy between two corresponding data elements. The flag setting criterion may be absolute (whenever there is a non-zero difference), or whenever there is a statistically significant difference. To conserve resources, a checksum scheme may be used instead on a parameter-by-parameter basis. Native and HDF-EOS data for each parameter would be summed and compared.

The native format read program and the swath read program may be readily modified to accomplish the aforementioned tasks. A backward migration may not be practicable. The data would have to be re-packed to do a volume check. The native format read program may not be readily modified to perform the re-packing.

As an example, we developed a simple validation scheme based on element by element checking and checksum procedures. The 22 satellite position parameters from the ERBE S-8 PAT product were used for the test case by comparing each native format data element with the corresponding HDF-EOS data element using a percent differencing algorithm. In addition, checksums for the two formats are compared. To seven digit accuracy, no errors were detected.

The C program used is listed below.

```
#include <math.h>
#include <stdio.h>

main()
{
    int sumerr,errflg[22];
    double nati,hdfi;
    double diff,checksum_nat,checksum_hdf;
    float pct_diff;
    int i,j;

    FILE *hdfin,*natin;

/* OPEN FILES */
    if((natin = fopen("nat.asc","r")) == NULL)
    {
        printf("Cannot open nat.asc\n");
        exit(1);
    }

    if((hdfin = fopen("hdf.asc.flawed","r")) == NULL)
    {
        printf("Cannot open hdf.asc\n");
        exit(1);
    }

/* PRINT TITLES */
    printf("\n\n\n");
    printf("Satellite Position Parameters\n\n");
    printf(" N  Native          HDF-EOS         ERR FLAG   %% DIFF\n");
    printf("\n");

/* ZERO CHECKSUMS */
    checksum_nat=0.;
    checksum_hdf=0.;
```

```
/* ZERO ERROR FLAGS */
    for(j=0;j<22;++j)
        errflg[j] = 0;

    for(j=0;j<22;++j)
    {
/* READ DATA */
        fscanf(natin, "%i %lf", &i, &nati);
        fscanf(hdfin, "%i %lf", &i, &hdfi);

/* IF THERE IS A DISCREPANCY; SET ERROR FLAG */
        if( (hdfi - nati) != 0.000)
            errflg[j] = 1;

/* FIND PCT DIFF */
        if(nati != 0.)
        {
            diff = hdfi - nati;
            pct_diff = fabs(diff) / fabs(nati) * 100. ;
        }
        else
        {
           pct_diff = -999.;
        }

/* PRINT DATA AND ERROR FLAGS*/
        printf("%2i %15.6f %15.6f      %i      %.6f\n",
            i,nati,hdfi,errflg[j],pct_diff);

/* FIND CHECKSUMS */
        checksum_nat += nati;
        checksum_hdf += hdfi;
    }
/* PRINT CHECKSUMS */
    if( (checksum_hdf - checksum_nat) == 0.000000)
    {
        sumerr = 0;
    }
    else
    {
        sumerr = 1;
    }

    printf("\nCHECKSUMS\n\n");
    printf("   %15.6f %15.6f      %i\n",checksum_nat,checksum_hdf,sumerr);
}
```

The following sample printout, generated by the above source code, shows a perfect one-to-one comparison of the 22 data elements.

Satellite Position Parameters

N	Native	HDF-EOS	ERR FLAG	% DIFF
1	2446431.000000	2446431.000000	0	0.0000000
2	0.500185	0.500185	0	0.0000000
3	0.983269	0.983269	0	0.0000000
4	4705621.000000	4705621.000000	0	0.0000000
5	4787850.000000	4787850.000000	0	0.0000000
6	-3477352.000000	-3477352.000000	0	0.0000000
7	-3448990.000000	-3448990.000000	0	0.0000000
8	-3814399.000000	-3814399.000000	0	0.0000000
9	-3736844.000000	-3736844.000000	0	0.0000000
10	5435.000000	5435.000000	0	0.0000000
11	5348.000000	5348.000000	0	0.0000000
12	2089.000000	2089.000000	0	0.0000000
13	2148.000000	2148.000000	0	0.0000000
14	4811.000000	4811.000000	0	0.0000000
15	4882.000000	4882.000000	0	0.0000000
16	123.099998	123.099998	0	0.0000000
17	122.349998	122.349998	0	0.0000000
18	323.539978	323.539978	0	0.0000000
19	324.229980	324.229980	0	0.0000000
20	113.050003	113.050003	0	0.0000000
21	180.750000	180.750000	0	0.0000000
22	6745.000000	6745.000000	0	0.0000000

CHECKSUMS

```
   -2505036.496589 -2505036.496589      0
```

The next sample output shows a comparison where some of the data elements were rounded off or by a hypothetical bug in the HDF-EOS migration.

Satellite Position Parameters

N	Native	HDF-EOS	ERR FLAG	% DIFF	
1	2446431.000000	2446431.000000	0	0.0000000	
2	0.500185	0.500185	0	0.0000000	
3	0.983269	0.983269	0	0.0000000	
4	4705621.000000	4705621.000000	0	0.0000000	
5	4787850.000000	4787850.000000	0	0.0000000	
6	-3477352.000000	-3477352.000000	0	0.0000000	
7	-3448990.000000	-3448990.000000	0	0.0000000	
8	-3814399.000000	-3456556.000000	1	9.3813734	***Bad Data***
9	-3736844.000000	-3736844.000000	0	0.0000000	
10	5435.000000	5435.00000000	0	0.0000000	
11	5348.000000	5348.000000	0	0.0000000	
12	2089.000000	2089.000000	0	0.0000000	
13	2148.000000	2150.000000	1	0.0931099	***Major round-off error***
14	4811.000000	4811.000000	0	0.0000000	
15	4882.000000	4882.000000	0	0.0000000	
16	123.099998	123.099998	0	0.0000000	
17	122.349998	122.350000	1	0.0000016	***Insignificant round-off error***
18	323.539978	323.539978	0	0.0000000	
19	324.229980	324.230000	1	0.0000062	***Insignificant round-off error***
20	113.050003	113.050000	1	0.0000027	***Insignificant round-off error***
21	180.750000	180.750000	0	0.0000000	
22	6745.000000	6745.000000	0	0.0000000	

CHECKSUMS

-2505036.496589	-2147191.496570	1

References

[1] Science Data Plan, Version 3, NASA, Goddard Space Flight Center, July, 1994.

[2] HDF-EOS Primer for Version 1 EOSDIS, EDHS[1] 175-WP-001-001, April, 1995.

[3] The HDF-EOS Swath Concept, EDHS[1], 170-WP-003-001, December, 1995.

[4] The HDF-EOS Point Concept White Paper, EDHS[2], Draft, ECS, February, 1996.

[5] The HDF-EOS Grid Concept White Paper, EDHS[2], Draft, ECS, February, 1996.

[6] Thoughts on HDF-EOS Metadata, EDHS[1], 170-WP-002-001, December, 1995.

[7] SDPS Database Design and Database Schema Specification, EDHS[1], DID311-CD-002-005, May, 1996.

[8] Software Metrics, R. B. Grady & D. L. Caswell, Prentice-Hall, Inc., 1987.

[9] The Art of Computer Systems Performance Analysis, R. Jain, John Wiley & Sons, Inc, 1991.

[10] Pilot Data Migration Report, EDHS[1], 160-TP-006-001, HAIS, October 30, 1995.

[11] Mathematical Statistics, J. E. Freund & R. E. Walpole, Prentice-Hall, Inc.,1987.

[12] Sample Survey Methods and Theory Volume I, M. H. Hansen, W. N. Hurwitz & W. G. Madow, John Wiley & Sons, Inc.,1953.

[13] Statistical Processes and Reliability Engineering, D.N.Chorafas, D. Van Nostrand Company Inc.,1960.

[14] Numerical Recipes in C: The Art of Scientific Computing, Cambridge University Press, 1992 or go to Numerical Recipes web site for sample articles; http://world.std.com/~nr

---

[1] ECS Data Handling System (EDHS) web site; http://edhs1.gsfc.nasa.gov

[2] HDF-EOS web site; http://edhs1.gsfc.nasa.gov/ftp/hdf_eos; ID = hdfeos; pswd = load2me